SEARCHING BEST FITTED REGRESSION MODELS FOR THE PRODUCTION OF MANGO AND SUGARCANE YIELDS

S.K. Yadav*, Diksha Arya*, Rishabh Kumar*, Taruna Taruna**, Carlos Bouza***
*Department of Statistics, School of Physical and Decision Sciences, Babasaheb Bhimrao Ambedkar University, Lucknow, India
**Department of Rural Management Babasaheb Bhimrao Ambedkar University, Lucknow, INDIA
***Facultad de Matemática Computación, Universidad de La Habana, Ciudad de La Habana, La Habana, Cuba

ABSTRACT
In this paper, we have compared some linear and nonlinear models for explaining and forecasting the productionof two different crops, Mango and the Sugarcane using area of fields as the auxiliary variable. The models under considerations are compared on the basis of different fitting measures such as, Coefficient of Determination (R^2),Residual Mean Square (s^2), Mean Absolute Error (MAE) and Akaike Information Criterion (A.I.C.). The two primary data sets have been collected, first for sugarcane production from the Sitapur district of Uttar Pradesh state in India and second for mango production from the Lucknow district of Uttar Pradesh state in India. The fitting measures are calculated for the collected primary data sets and the best fitted models are selected and recommended for further use on the basis of the fitting measures. From the data analysis for both the data sets, it is found that the Compound, Growth, Exponential and Logistic models are equally good for practical applications.

KEYWORDS: Study Variable, Auxiliary Variable,Model Fitting, Fitting Measures, Linear and Non-linear Models.

MSC: 62P10

RESUMEN
En este paper, hemos comparado algunos modelos lineales y no-lineales para explicar y predecir la producción de dos diferentes cultivos, Mango y Caña de Azúcar usando el área de los campos como variable auxiliar. Los modelos bajo consideración son comparados en base a diferentes medidas de ajuste como, Coeficiente de Determinación (R^2), Cuadrado Medio Residual Mean (s^2), Error Absoluto Medio (MAE) y el Criterio de Información de Akaike (A.I.C.). Los dos primarias data-sets han sido colectadas. Primero para la producción de caña de azúcar del distrito de Sitapur del estado de Uttar Pradesh en India y el segundo para la de mango para el distrito de Lucknow del estado de Uttar Pradesh, India. Las medidas de ajuste fueron calculadas para los datos primarios y los segundos mejores modelos son seleccionados y recomendados a partir de las medidas de ajuste. Del análisis de la de los conjuntos de datos se halló que fueron igualmente buenos para la aplicación los modelos Compound, Growth, Exponential and Logistic.

PALABRAS CLAVE: Variable de estudy, Variable Auxiliar, Ajuste de Modelos, Medidas de Ajuste, Modelos Lineale y Non-lineales

1. INTRODUCTION

The term regression was first used as a statistical concept in late 18th century. The word “regress” signified “moving back to the average”. Nowadays, regression analysis is a collection of several methods which deals with establishing the functional relationship between a dependent variable and one or several independent variables. More specifically regression analysis is a statistical technique to predict one variable from another variable. The variable about which we have full information is known as independent variable and the main unknown variable under consideration is called as dependent variable. For instance, if there is a linear relationship between the age of father and son then if we know the age of any one of them then we can get the age of another by using that relationship. The method of least squares, proposed by Adrien-Marie-Legendre in 1805. See if interested in its roots Legendre (1805):, and Carl Fredrich Gauss (1809). Legendre suggested using the least square approach, which minimises the sum of squares of these variations. Carl Friedrich Gauss, a German mathematician who may have used the same method before, made significant computational and theoretical breakthroughs. In 1821, Gauss published a continuation of the theory of least squares, which included a variant of the Gauss Markov Theorem. You may consult this oeuvre in a recently publication of it, Gauss (1995). Initially , Francis Galton developed the term regression to describe a biological process in the nineteenth century. In this phenomenon he observed that the heights of descendants (an ancestor refers to any element which is connected further up no matter how many levels higher) tend to regress down towards a normal average. The phenomenon, was also known as regression towards the mean. Regression had just this biological meaning for Galton, but his work was eventually extended to a broader statistical context by Udny Yule and Karl Pearson. The joint distribution of the main and independent variables is assumed to be Gaussian in Yule and Pearson's works. R.A. Fisher's efforts from 1922 and 1925 undercut this idea. The conditional distribution of the main variable was assumed to be Gaussian by Fisher, but the joint distribution did not have to be. Fisher's assumption is similar to the Gauss Formulation of 1821 in this regard. Regression analysis is most typically used to calculate the dependent variable's conditional expectation given the independent variables, or the average value of the dependent variable while the independent variables are fixed. For prediction and forecasting, regression analysis is commonly utilised. It's also used to figure out which of the independent factors are linked to the dependent variables, as well as to investigate the nature of those links. Interested in historical details see Galton (1877):

Resuming, Linear Regression is a linear model with a linear connection between the input variables (X) and the single output variable (Y). As a result, a linear combination of the input variables can be determined (X). The method is known as a simple linear regression model when there is only one input variable (X). When there are several input variables, the procedure is referred to as multiple linear regression in statistics literature. Simple linear regression relates two variables (Y and X) with a straight line ($Y = a + bX$), while nonlinear regression relates the two variables in a nonlinear relationship. Normality, linearity, homoscedasticity, multicollinearity and autocorrelation are some of the assumptions to proceed for regression analysis. Normality means how well the data is modelled by a normal distribution. Statisticians utilise the Histogram, Normal Probability Plot, and Scatterplot of Residuals to check the data's normality. After creating a histogram, we can look at the line that displays the distribution's shape to see if it is genuinely normal or not. Is expected that on the Normal Probability Plot the actual observations are falling near to the diagonal from lower left to upper right. If the data has a normal distribution, the residuals scatter plot reveals that the majority of the residuals for the corresponding value of the predicted score are in the centre of the plot, with some residuals trailing out symmetrically from the centre.

The assumption of homoscedasticity states that the residuals for all projected dependent variables are approximately equal. We can check for homoscedasticity by looking at a scatter plot between each independent variable and dependent variable. Singularity occurs when the independent variables are perfectly correlated and one independent variable is a combination of one or more of the other independent variables. Multicollinearity occurs when the independent variables are highly correlated (0.90 or greater), whereas singularity occurs when the independent variables are perfectly correlated and one independent variable is a combination of one or more of the other independent variables. Strong bivariate correlations or high multivariate correlations can create multicollinearity and singularity. Simple running correlations among independent variables can easily reveal high bivariate correlations. We are only looking at the scenario of one independent variable in this study. See generalities and discussions in Aldrich (2005), the roots in Pearson et al. (1903):

2. MATERIALS AND METHODS

It is well known that the general linear regression model with Y as the dependent variable and Y as the independent variable is presented as:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Where, Y is termed as the dependent or study variable and X is termed as independent or explanatory variable. The terms β_0 and β_1 are the parameters of the model. The parameter β_0 is termed as intercept term and the parameter β_1 is termed as slope parameter. These parameters are usually called as regression coefficients. The unobservable error component ε accounts for the failure of data to lie on the straight line and represents the difference between the true and observed realization of Y. Its modern development is due to the seminal works discussed in Fisher (1922) and presented in the text of (1954)..

The Ordinary Least Square (OLS) Estimates of β_0 and β_1 are,

$$\hat{\beta}_0 = \hat{Y} - \hat{\beta}_1 \overline{X}$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum {x_i}^2}$$

Where, $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$. Theoretical discussions are given in standard text books as Montgomery, Peck &Vining(2012), Rencher & Schaalje (2008).

The mathematical form of the Nonlinear Regression model is given by,

$$\mathbf{Y} = \mathbf{f}(\mathbf{x}, \boldsymbol{\theta}) + \boldsymbol{\varepsilon}$$

Where, θ is vector of unknown parameters and ε is an uncorrelated random error term. We also typically assume that the errors are normally distributed, as in the case of linear regression model. The general solution is obtainable using Mathematical Programming methods, see as examples Allende & Bouza

(1995, 1999) and Allende et al (2003). Commonly the parameters of the nonlinear regression models are estimated by the method of Nonlinear Least Square, Method of Three Selected Points, Steepest Descent Method and the well-known Levenberg Marquardt methods. See details books as Gujarati & Sangeetha (2008), Fox (2016). Among these methods, Levenberg Marquardt is the best one and we have used it for the estimation of the parameters of the nonlinear models.

There are different fitting measures for the regression models and on the basis of these measure, we find the best fitted models for the better explanation and future prediction of the phenomenon under consideration. Following are different goodness of fit measures:

Coefficient of Determination (R^2)

The regression model is evaluated by determining how much of the overall sum of squares has dropped into the sum of squares as a result of the regression. The coefficient of determination R^2 is defined as,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

The R^2 value closer to 1 describe that the most of the amount of variability in Y has been explained by the fitted model. Thus, observance of a high R^2 is an indication of the good fit.

Coefficient of Determination -$\mathbf{R^2_{Adj}}$

The measure R^2_{Adj} was described by Montgomery *et al.* (2012) as being good for model comparison when the numbers of parameters in two models are not equal. The coefficient of determination $R^2_{Adj}$is defined as,

$$R^2_{Adj} = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2/n-p}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2/n-1}$$

The value of R^2_{Adj}reaching very close to 1 implies that most of the variability in the data has been explained by the fitted model. Thus, observance of a high value indicates a good fit.

Residual Mean Square (s^2)

The residual mean square s^2 for the models for our problem is defined as,

$$s^2 = \frac{\sum_{i=1}^{n}\left(S_w^2 - \widehat{S_w^2}\right)^2}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p}$$

Where, the number of observations is n, while the number of model parameters utilised is p. A small s^2 number indicates that the error-related regression is small, implying that the sum of squares owing to regression is high enough to match the entire sum of squares. Hence a s^2 value indicates that the fitted model is adequate.

Mean Absolute Error (MAE)

The MAE which is average of absolute error is defined as,

$$MAE = \frac{\sum_{i=1}^{n}\left|Y_i - \widehat{Y}_i\right|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n}$$

Where, n is the number of observations. A smaller MAE is preferred in fitting of various linear and nonlinear regression models.

Akaike Information Criterion (A.I.C.)

The Akaike Information Criterion (A.I.C.) was given a lot of weight , see Gujarati and Sangeetha (2007). It is a very useful criterion for judging the performance of fitted model. It is also good for comparing two or more models. The model with lowest value of A.I.C. is preferred. It can be defined as,

$$\text{A.I.C.} = \exp\left(\frac{2p}{n}\right)\frac{RSS}{n}$$

Where, the number of observations is n, while the number of parameters is p. RSS stands for residual sum of squares.

3. MODELS UNDER CONSIDERATION

The purpose of the present study is to find the best fitted regression model for the explanation and prediction of the production of the Sugarcane and the Mango yields at two different district of Uttar Pradesh State in India. Following are different linear and nonlinear regression models under comparison presented in Table-1 below as:

Table1: Different linear and nonlinear models under consideration

Model	Functional Form
Linear	$Y = \beta_0 + \beta_1 x$
Logarithmic	$Y = \beta_0 + \beta_1 \ln(x)$
Inverse	$Y = \beta_0 + (\beta_1 / x)$

Quadratic	$Y = \beta_0 + (\beta_1 x) + (\beta_2 x^2)$
Cubic	$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$
Power	$Y = \beta_0 (x^{\beta_1})$
Compound	$Y = \beta_0 (\beta_1{}^{x})$
S-curve	$Y = \exp(\beta_0 + (\beta_1/x))$
Logistic	$Y = 1 / (1/u + (\beta_0 (\beta_1{}^{x})))$
Growth	$Y = \exp(\beta_0 + (\beta_1 x))$
Exponential	$Y = \beta_0(\exp(\beta_1 x))$

We have collected two primary data sets and a simulated data for the numerical comparison of different models under consideration.

Data Description of Data-1

We have collected primary data of Sugarcane Production from some villages of Machrehta block of Sitapur District for 100 fields. The details of the collected data are as follows:

Dependent Variable: Yield of Sugarcane denoted as Y in quintals (1 quintal=100 Kilogram)

Independent Variable: Area of cultivation denoted as X in Bigha (1 Bigha=0.2508 hectare)

Descriptive Statistic: $N = 100$, $\bar{X} = 8.0275$, $\bar{Y} = 468.745$, $\sigma_X = 5.598396122$, σ_Y =365.8315386, $\rho_{XY} = 0.968655062$

The data collected from 100 fields is presented in Table-2 below.

Table-2: Data-1 collected on sugarcane yield form Sitapur District

S.N.	Y	X	**S.N.**	Y	X	**S.N.**	Y	X	**S.N.**	Y	X
1	900	18	26	125	2	51	119	2	76	450	10
2	2450	35	27	900	17	52	180	4	77	350	6
3	150	4	28	650	12	53	64.5	1	78	1000	20
4	500	8	29	1300	20	54	160	3	79	500	9
5	450	10	30	90	2	55	128	2	80	350	6.25
6	150	5	31	170	3	56	490	7	81	700	10
7	125	3	32	225	4	57	350	7	82	750	12
8	250	4.25	33	300	5	58	230	4	83	700	12
9	550	8.5	34	80	2	59	550	10	84	400	9
10	200	2	35	400	7	60	520	8	85	300	4
11	200	4	36	200	4	61	450	8	86	500	8
12	400	7	37	900	12	62	850	12	87	150	3
13	250	4	38	150	3.5	63	650	10	88	350	5
14	900	15	39	800	14.5	64	950	12	89	250	4
15	350	5	40	281	6	65	750	13.5	90	150	3
16	850	12	41	158	4	66	600	9	91	250	5
17	450	10.5	42	900	15	67	189	3	92	300	7
18	1000	15	43	90	1.5	68	450	7	93	500	8
19	300	4.75	44	400	7.5	69	500	11	94	100	2
20	100	3.5	45	650	10	70	1000	20	95	400	6
21	200	3	46	800	13	71	850	12	96	450	7
22	200	3.5	47	350	6	72	800	12	97	200	3
23	850	15	48	350	6	73	500	9	98	300	6
24	150	3	49	1750	25	74	650	15	99	100	2
25	350	6	50	350	7	75	300	6	100	450	10

The following graph, Figure-1, represents fitting of different considered linear and non-linear regression models to the data-1.

Figure-1: Fitting of different linear and non-linear models to data-1

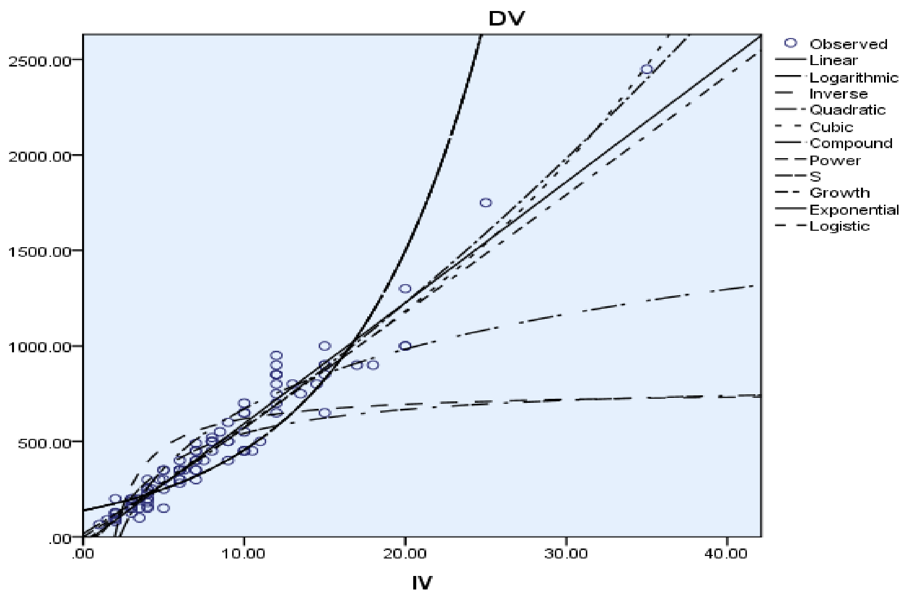
The various fitting measures for the linear and non-linear regression models under consideration are presented in Table-3 below.

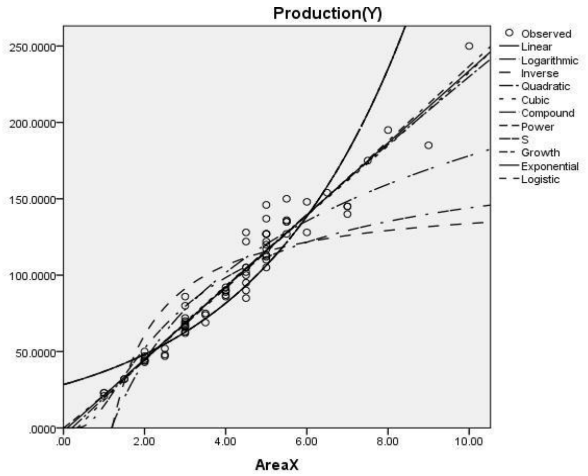
Table-3: Fitting measures of different linear and non-linear regression models

Model	R^2	s^2	**MAE**	**AIC**
Linear	0.938	817588.018	65.2570	8509.54
Logarithmic	0.738	3475896.782	474.8450	36177.51
Inverse	0.413	7780742.93	187.1220	80482.81
Quadratic	0.945	726017.542	61.5440	7709.12
Cubic	0.947	698237.382	60.6250	7563.92
Compound	0.796	11.327	172.8010	0.11789
Power	**0.933**	**3.716**	**64.6817**	**0.03867**
S	0.743	14.253	131.0460	0.14834
Growth	0.796	11.327	170.8800	0.11789
Exponential	0.796	11.327	170.8900	0.11789
Logistic	0.796	11.327	437.1720	0.11789

Data Description of Data-2

A primary data on Mango





Production has been collected from the Kakori Block of Lucknow District at Uttar Pradesh state in India. The data has been collected on 72 mango orchards from different villages of Kakori Block. The data description is given as:

Dependent Variable: Production denoted as Y in quintals

Independent Variable: Area of cultivation denoted as X in Bigha

Descriptive Statistic: $\mathrm{N} = 72, \bar{X} = 4.180555556,$ $\bar{Y} = 96.45833333, \sigma_X = 1.829541533,$ $\sigma_Y = 44.40908976, \rho_{XY} = 0.972248893$

The data collected from 72 mango orchards is presented in Table-4 below.

Table-4: Data-2 collected on mango yield form Lucknow District

S.N.	Y	X	S.N.	Y	X	S.N.	Y	X	S.N.	Y	X
1	110	5	19	23	1	37	135	5.5	55	154	6.5
2	195	8	20	117	5	38	67	3	56	105	4.5
3	100	4.5	21	45	2	39	86	4	57	86	3
4	145	7	22	74	3.5	40	21	1	58	140	7
5	136	5.5	23	68	3	41	75	3.5	59	120	5
6	146	5	24	66	3	42	66	3	60	90	4
7	70	3	25	113	5	43	47	2.5	61	105	4.5
8	90	4.5	26	127	5.5	44	32	1.5	62	150	5.5
9	145	7	27	92	4	45	63	3	63	80	3
10	50	2	28	250	10	46	127	5	64	122	5
11	128	6	29	137	5	47	95	4.5	65	175	7.5
12	148	6	30	69	3	48	52	2.5	66	122	4.5
13	90	4	31	87	4	49	127	5	67	128	4.5
14	62	3	32	63	3	50	48	2.5	68	105	5
15	44	2	33	69	3.5	51	43	2	69	102	4.5
16	135	5.5	34	89	4	52	185	9	70	44	2
17	67	3	35	47	2	53	72	3	71	23	1
18	127	5	36	32	1.5	54	112	5	72	85	4.5

The following graph Figure-2 represents fitting of different considered linear and non-linear regression models to the data-2.

Table-5: Fitting measures of different linear and non-linear models to data-2

Model	R^2	s^2	MAE	AIC
Linear	0.945	109.483	7.408	112.522
Logarithmic	0.856	288.519	12.728	296.53
Inverse	0.608	784.984	21.502	806.778
Quadratic	0.946	110.384	7.475	111.828
Cubic	0.946	111.542	7.648	111.363
Compound	0.865	0.037	15.230	0.0381
Power	**0.967**	**0.009**	**7.285**	**0.009**
S	0.857	0.039	14.850	0.040
Growth	0.865	0.037	15.230	0.038
Exponential	0.865	0.037	15.230	0.038
Logistic	0.865	0.037	15.230	0.038

. **Figure-2: Fitting of different linear and non-linear models to data-2 regression models**

The various fitting measures for the linear and non-linear fitted regression models under consideration are presented in Table-5.

Simulation Study

This section presents the procedure and analysis for the two simulated population each considering the parameters of the first and second primary data. It is required to generate a bivariate population with a specified correlation. We have used the mean, standard deviation and correlation coefficient of the primary data to simulate the population of 10,000 observations.

Data Description of Simulated Data-1

To generate simulated population, we have considered the same parameters of the two primary data. The simulated population is generated through bivariate normal distribution with their mean vector and variance-covariance matrix as:

Means of [Y, X] as$\mu = [468.745, 8.0275]$.

Variances and covariance of [Y, X] as $\sigma^2 = \begin{bmatrix} 31.34204 & 1983.873 \\ 1983.87324 & 133832.715 \end{bmatrix}$.

With correlation$\rho_{YX} = 0.968655062$

Hence a bivariate normal distribution of X and Y of size N = 10,000 have been generated through these parameters using R Program.

The following graph Figure-3 represents fitting of different considered linear and non-linear regression models to the simulated data-1.

Table-6: Fitting measures of different linear and non-linear regression models

Model	R^2	s^2	MAE	AIC
Linear	0.919	8178.74	72.02	8180.378
Logarithmic	0.626	37620.95	147.61	37628.48
Inverse	0.011	99469.02	258.38	99488.92
Quadratic	0.920	7997.99	71.36	7998.795
Cubic	0.921	7925.23	71.12	7925.237
Compound	0.634	0.369	158.25	0.369282
Power	**0.661**	**0.343**	**92.12**	**0.342636**
S	0.026	0.983	266.56	0.983652
Growth	0.634	0.369	158.25	0.369282
Exponential	0.634	0.369	158.25	0.369282
Logistic	0.634	0.369	158.25	0.369282

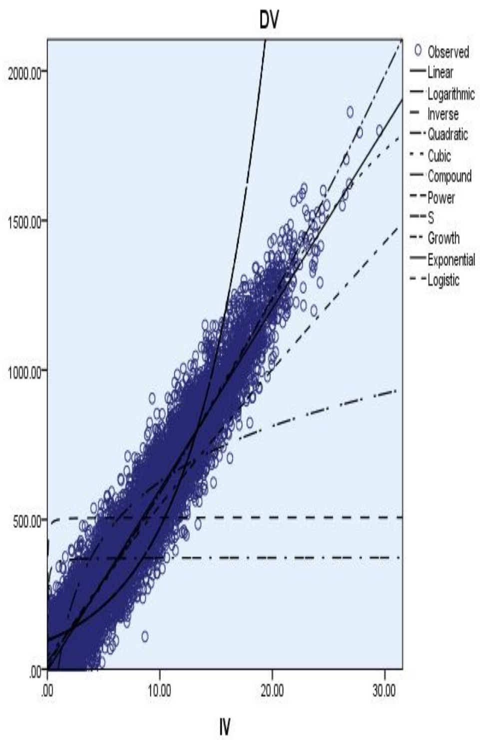


Figure-3: Fitting of different linear and non-linear models to simulated data-1

The various fitting measures for the linear and non-linear fitted regression models under consideration are presented in Table-6.

Data Description of Simulated Data-2

To generate simulated population, we have considered the same parameters of the two primary data. The simulated population is generated through bivariate normal distribution with their mean vector and variance-covariance matrix as:

Means of [Y, X] as$\mu = [96.45833333, 4.180555556]$.

Variances and covariance of [Y, X] as $\sigma^2 = \begin{bmatrix} 3.347222 & 78.99354 \\ 78.993545 & 1972.16725 \end{bmatrix}$.

With correlation $\rho_{YX} = 0.972248893$

Hence a bivariate normal distribution of X and Y of size N = 10,000 have been generated through these parameters using R Program.

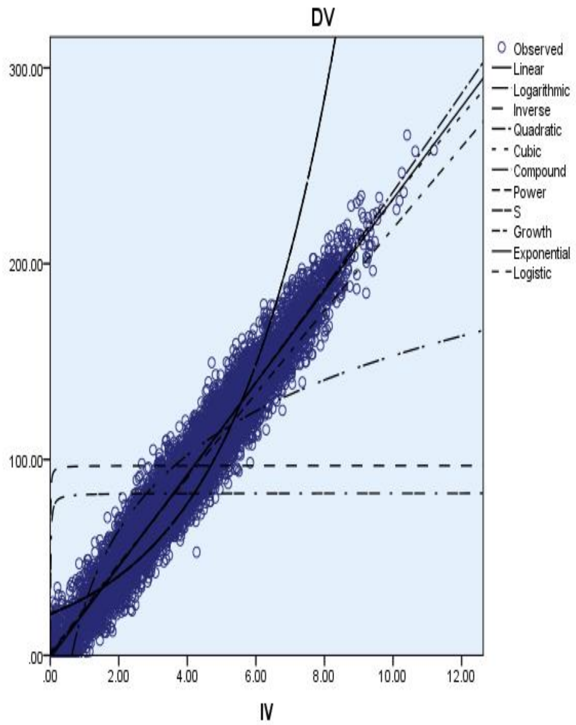
The following graph Figure-4 represents fitting of different considered linear and non-linear regression models to the simulated data-2.
The various fitting measures for the linear and non-linear fitted regression models under consideration are presented in Table-7 below.

Table-7: Fitting measures of different linear and non-linear regression model

Model	R^2	s^2	MAE	AIC
Linear	0.945	105.973	8.2196	105.9941
Logarithmic	0.699	577.031	17.3478	577.1465
Inverse	0.005	1905.22	35.1740	1905.608
Quadratic	0.945	105.712	8.2105	105.7221
Cubic	0.945	105.606	8.2045	105.6055
Compound	0.710	0.142	18.1325	0.141961
Power	**0.818**	**0.089**	**8.8283**	**0.088901**
S	0.017	0.481	36.6403	0.481289
Growth	0.710	0.142	18.1325	0.141961
Exponential	0.710	0.142	18.1325	0.141961
Logistic	0.710	0.142	18.1325	0.141961

Figure-4: Fitting of different linear and non-linear models to simulated data-2

4. RESULTS AND DISCUSSION



Following are some of very interesting observations to be considered:

1. From Table-3, it may be observed that the "Power" non-linear model is the best fitted model to the data set-1 of the sugarcane yield. It is the best fitted model for predicting sugarcane yield as compared to the considered linear and nonlinear models because it have least values of s^2, MAE and AIC fitting measures.

2. From Table-5, it is evident that the "Power" model is the best fitted model for the prediction of mango production in comparison to linear and non-linear competing estimators and this model have smaller values of s^2, MAE and AIC fitting measures. The values of R^2 are not compared as it may give misleading information in case of non-linear models.

3. From Table-6, it may easily be seen that the "Power" model is the best fitted model for the simulated data on sugarcane production using the same parameters of the real data set-1 among the class of all competing linear and non-linear models as the fitting measures obtained the optimum values for this model.

4. From Table-7, it may be verified that the “Power” model is the best fitted model for the simulated data on mango production using the same parameters of the real data set-2 among the class of all competing linear and non-linear models as the fitting measures are having the optimum values for this model.

5. CONCLUSIONS

In this paper, we have fitted eleven different linear and non-linear models namely Linear, Logarithmic, Inverse, Quadratic, Cubic, Compound, Power, S, Growth, Exponential and Logistic for the analysis and prediction of Sugarcane and Mango productions at Sitapur and Lucknow districts of Uttar Pradesh State in India to two real primary data sets. The two simulated populations each of size 10000 for the same parameters of the real data sets with the bivariate normal distribution are generated and these models are also fitted to these simulated populations. From Table-3, Table-5, Table-6 and Table-7, it may easily be observed that the “Power” non-linear model is the best for all real and simulated populations. It exhibited the smallest values of the residual sum of squares s^2, which is one of the most suitable fitting measure among other fitting measures. Thus, the “Power” non-linear model may be used for the better explanation and future predictions of the Sugarcane and Mango productions and different policies may be formed accordingly.

Acknowledgements: C. N. Bouza acknowledges the support of CITMA project **PN223LH010-033.**

REFERENCES

[1] ALDRICH, JOHN (2005): Fisher and Regression. **Statistical Science**. 20, 401 - 417. doi:10.1214/088342305000000331.

[2] ALLENDE ALONSO, S. M., C. N. BOUZA HERRERA (1995): Multiple linear regression curve fitting: a quadratic programming solution. In “**Approximation & Optimization**” P. Lang Verlag, Frankfurt. 1-10

[3] ALLENDE ALONSO, S. M., C. N. BOUZA HERRERA (1999): El Modelo de regresión lineal : algunas soluciones para su ajuste. **Inv. Operacional**, 20 115-40.

[4] ALLENDE ALONSO, S. M., C. N. BOUZA HERRERA AND GEMAYQZEL BOUZA ALLENDE (2003): El uso de recocido simulado para la selección del mejor modelo de regresión: el caso L_2**. Inv. Operacional**, 24, 292-299.

[5] FISHER, R.A. (1922): The goodness of fit of regression formulae, and the distribution of regression coefficients**. Journal of the Royal Statistical Society**. 85, 597-612.

[6] FISHER, R.A. (1954): **Statistical Methods for Research Workers** (Twelfth ed.). Edinburgh: Oliver and Boyd. ISBN 978-0-05-002170-5.

[7] FOX, R, (2016): **Applied regression analysis and generalized linear models,** (3rd ed.), Dage Publications Inc, Thousand Oaks, CA

[8] GALTON F. (1877): Typical laws of heredity, **Nature** 15, 492–495, 512–514, 532–533.

[9] GAUSS, C. F. (1995): **Theory of the Combination of Observations Least Subject to Errors,(original and translation of Theoria Combination is Observation um Erroribus Minimis Obnoxiae**, Society for Industrial and Applied Mathematics, Philadelphia, PA.

[10] GUJARATI, D. and SANGEETHA, N. (2007): **Basic Econometrics**. Fourth Edition, Tata McGraw-Hill, New Delhi.

[11] LEGENDRE, A.-M. (1805): **Nouvelles méthodes pour la détermination des orbites des comètes** [*New Methods for the Determination of the Orbits of Comets*] (in French), Paris: F. Didot, hdl:2027/nyp.33433069112559.

[12] HOSMER, D. W., JR., & LEMESHOW, S. (2000): **Applied logistic regression** (2nd ed.). Wiley, N. York.

[13] MCCULLAGH, P., NELDER, J. (1989): **Generalized Linear Models**. 2nd Ed. Chapman & Hall, London.

[14] MONTGOMERY D. C., PECK E. A., VINING G. C. (2012): **Introduction to Linear Regression Analysis,** 5th Ed., Wiley, N. York.

[15] PEARSON, K. YULE, G.U. BLANCHARD, N. LEE, A. (1903): The Law of Ancestral Heredity. **Biometrika.** 2 ,211-236. doi:10.1093/biomet/2.2.211. JSTOR 2331683.

[16] QUEIROZ, F. F. ,and S. L. P. FERRARI (2022): Power logit regression for modeling bounded data. **arXiv:2202.01697v1**

[17] R CORE TEAM (2021): **R: A Language and Environment for Statistical Computing. R Foundation**

for Statistical Computing. Vienna, Austria.

[18] RENCHEr A. C. and G. B. Schaalje (2008): **Linear Models in Statistics**, 2nd edition, Wiley, Hoboken, New Jersey.

[19] YULE, G. UDNY (1897): "On the Theory of Correlation". **Journal of the Royal Statistical Society**. 60, 812–54. doi:10.2307/2979746. JSTOR 2979746.

[20] SRINIVAS, J., BOIROJU, N.K., REDDY, M.K. (2013): Comparison of Ratio Estimators of Population Mean, **International Journal of Mathematical Sciences, Technology and Humanities** 108, 1158–1163